CAFA: a Controllable Automatic Foley Artist

Anonymous ICCV submission

Paper ID 5746

Abstract

Foley is a key element in video production, refers to the process of adding an audio signal to a silent video while ensuring semantic and temporal alignment. In recent years, the rise of personalized content creation and advancements in automatic video-to-audio models have increased the demand for greater user control in the process. One possible approach is to incorporate text to guide audio generation. While supported by existing methods, challenges remain in ensuring compatibility between modalities, particularly when the text introduces additional information or contradicts the sounds naturally inferred from the visuals. In this work, we introduce CAFA *(Controllable Automatic Foley Artist) a video-and-text-to-audio model that generates semantically and temporally aligned audio for a given video, guided by text input.* CAFA *is built upon a text-to-audio model and integrates video information through a modality adapter mechanism. By incorporating text, users can refine semantic details and introduce creative variations, guiding the audio synthesis beyond the expected video contextual cues. Experiments show that besides its superior quality in terms of semantic alignment and audio-visual synchronization the proposed method enable high textual controllability as demonstrated in subjective and objective evaluations.* [1]

1. Introduction

In recent years, personal content creation has become a major part of everyday life, shaping how we work, entertain, and communicate. One example is *Foley*, the art of adding sound effects to silent videos while ensuring precise semantic and temporal alignment [1]. Traditionally, this process was done manually by professional sound designers. However, with the growing demand for fast and immediate personal digital content, the need for automation and accessibility of this process has increased. An effective Foley generation approach should produce high-quality, synchronized audio while also allowing users to creatively shape the sound, balancing precision with creative flexibility.

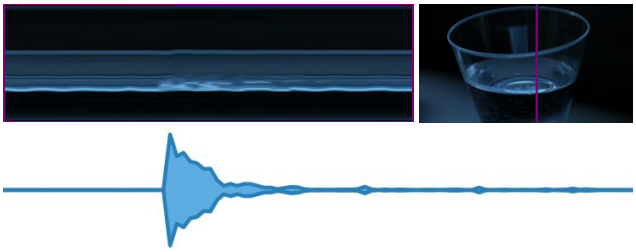

Figure 1. **Motivation.** An iconic scene from *Jurassic Park*, where water in a glass shakes due to the approaching footsteps of a T-Rex. Inferring the generated sound from the video alone is insufficient, as the task is inherently ambiguous. **Top:** a representative frame and a Y-T slice (from the purple column), where the temporal cue of the shake is faintly visible. **Bottom:** Our method leverages the prompt "T-Rex Stomping" to generate a synchronized audio track that aligns with both the visual timing and artistic intent.

Building on this need, recent advancements in generative models have led to the development of Video-to-Audio (V2A) models, which aim to automate Foley synthesis and explore cross-modal correspondences [17, 31, 43, 46, 48, 50]. While these models effectively capture semantic information through global visual representations such as CLIP [36], they often rely on motion-sound relationships for temporal alignment [31, 46, 48]. Some approaches, including [17], model motion explicitly using optical flow [14], while others [31, 46, 48] leverage contrastive learning-based encoders such as CAVP [31] and AV-CLIP [19] to learn temporally and semantically aligned audio-visual features. Despite these advancements, existing models remain limited to extracting information from the video itself and struggle to incorporate user-provided cues, restricting flexibility and creative control over sound design.

To bridge this gap, Text-and-Video-To-Audio (TV2A) models have been introduced, integrating textual information to enhance control over audio generation [5, 7, 20, 28, 50, 55]. By incorporating text, these models allow users to modify audio semantics, add details, and generate diverse sound variations. For instance, text can specify how

[1] Samples and code can be found in our anonymized demo page

a sound should be perceived, such as describing a door as creaking or coffee being sipped loudly. Another possibility is introducing creativity through text; a barking dog in a video could instead sound like a meowing cat or a crowing rooster, depending on the accompanying description. In the context of soundtrack design, one would often like to add sounds which do not appear in the video, such as in the iconic scene from *Jurassic Park*, where water in a glass shakes due to the approaching footsteps of a T-Rex; see Figure 1 for a visual example. However, textual conditioning is often not sufficiently strong or may come at the expense of temporal alignment between video and audio. Additionally, when the text describes semantics that differ from the video, existing models frequently struggle to generate a natural and coherent audio signal (See Section 5).

Various methods have been explored for integrating text into multimodal systems. A common strategy involves jointly training video, text, and audio representations to capture shared semantics [5, 7]. However, this requires retraining the entire network whenever modifications are made, leading to high computational costs. Alternatively, a training-free method [50] leverages a shared latent space to link the modalities, eliminating the need for retraining. Yet, this introduces test-time optimization, increasing inference time and potentially degrading output quality and alignment. A middle-ground solution employs a modality adapter (e.g., ControlNet mechanism [54]), which uses video inputs to condition a pretrained Text-to-Audio (T2A) model [20, 55], providing an effective way to incorporate video information into text-driven audio synthesis.

In this work, we introduce CAFA, which stands for Controllable Automatic Foley Artist, a novel text-and-video-to-audio model that extends beyond temporal and semantic synchronization, allowing users to shape and control sound through textual cues. CAFA leverages a ControlNet like modality adapter to flexibly integrate pretrained T2A models with video-based features while maintaining a relatively low training cost (48 A100 GPU hours for CAFA vs. 304 H100 GPU hours for the baseline method). Specifically, we explore Stable-Audio-Open [9] and TangoFlux [16] as T2A models. To extract temporal and semantic features, we experiment with AVCLIP [19] and CLIP [36] as the video representations. CAFA achieves high-quality audio synthesis, temporal synchronization, and contextual alignment performance comparable to state-of-the-art V2A and TV2A models. Additionally, it significantly surpasses existing TV2A approaches when the text and visual conditioning cues are semantically different, demonstrating greater control over generated sound.

Our main contributions are: (i) We introduce CAFA, a novel TV2A model that allows the generation of temporally and semantically aligned audio while providing extensive textual control over the generated audio; (ii) We evaluate CAFA against existing V2A and TV2A models, demonstrating comparable performance in audio quality and video-audio compatibility, while achieving superior performance for textual control, as validated through disentanglement experiments, objective evaluations, and human studies; (iii) CAFA is built on the modality adaptation (via a ControlNet mechanisem), enabling precise temporal control while offering a versatile framework that supports modular integration, accommodating different T2A models (Stable Audio Open and TangoFlux). Additionally, it facilitates the efficient incorporation of video representations, leading to more effective training compared to alternative methods.

2. Background

2.1. Latent Diffusion Models

Latent Diffusion Models (LDMs) [39] are a class of generative models that perform a diffusion process within a learned latent space $\mathbf{z}$. A Variational Autoencoder (VAE) encodes a data sample $\mathbf{x} \sim p(\mathbf{x})$ into a lower-dimensional latent space $\mathbf{z} \in \mathbb{R}^d$ using an encoder $\mathcal{E}$, while a decoder $\mathcal{D}$ reconstructs $\mathbf{x}$. Performing diffusion in this reduced space significantly reduces computational cost while maintaining high-quality generation.

The diffusion process follows two Markovian paths: the forward and reverse processes. In the forward process, a clean latent representation $\mathbf{z}_0$ is gradually corrupted with additive Gaussian noise:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \tag{1}$$

where $\{\alpha_t\}_{t=1}^T$ defines the noise schedule, $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A key consequence of the forward process is its marginal distribution:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{t=1}^T \alpha_t$. A neural network, trained as a denoiser, learns to estimate ϵ given the noisy input $\mathbf{z}_t$, the timestep t , and conditioning information c, such as encoded text. The training objective minimizes the difference between the true noise and the predicted noise:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t,c}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c)\|\right]. \tag{3}$$

By using the network output, the reverse process aims to reconstructs $\mathbf{z}_0$ from $\mathbf{z}_T$ by iteratively denoising it. While initial works [13, 44] formulated this process discretely, Song et al. [45] showed that it can be equivalently expressed as an Ordinary Differential Equation (ODE) which can be solved numerically using dedicated solvers. Specifically, Stable Audio Open [9] employs DPM-Solver++ [30] and follows the the v-objective approach [41].

Classifier-Free-Guidance (CFG) is a widely used method to improve performance in conditional generative models,

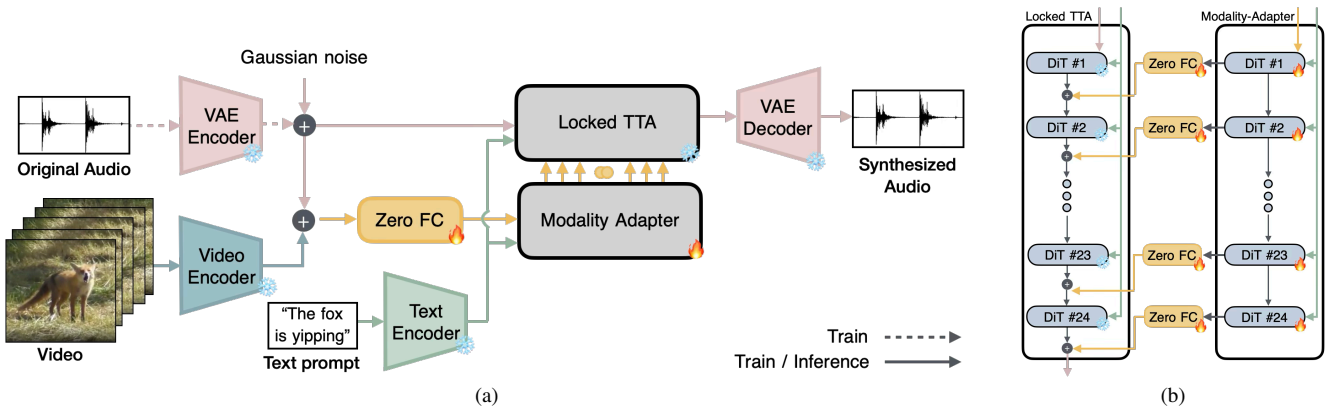

Figure 2. (a) **Method overview:** our model is text-and-video-to-audio, leverages pretrained models for audio generation, and video encoding. The original audio and VAE audio encoder are only used during training. (b) **Adaptor:** Illustration of the internal connectivity between the backbone T2A model and our video conditioning adaptor, with fully connected (FC) layers explicitly shown.

originally demonstrated in diffusion-based image generation approaches [12]. CFG is an effective control mechanism for steering the inference process to better align with provided conditioning signals. Specifically, it modifies the predicted noise by linearly combining the estimates from a conditional diffusion model and a jointly trained unconditional model, resulting in the following formulation:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, c, t) = \epsilon_\theta(\mathbf{z}_t, t) + \gamma\left(\epsilon_\theta(\mathbf{z}_t, t, c) - \epsilon_\theta(\mathbf{z}_t, t)\right) \quad (4)$$

where γ determines the strength of the guidance, with higher values enforcing stronger adherence to the conditioning signal.

2.2. ControlNet Mechanism

The ControlNet mechanism was initially introduced as a neural network architecture for controlling text-to-image models through spatially localized conditioning (e.g., canny edge and depth maps)[54]. It preserves the quality and stability of a large pretrained model by locking its weights, while enabling the incorporation of control signals through a replicated copy of that backbone model. These components are connected via zero-initialized convolutional layers, allowing a gradual integration which minimizes noise interference during training.

3. Method

CAFA consists of two main components: a pretrained backbone Text-To-Audio (T2A) model, with frozen weights to maintain audio quality, and a trainable modality adapter that integrates temporal and semantic video information. The components are linked through Zero Fully-Connected (FC) Layers, which prevent noise from disrupting the backbone model during early training. This structure allows us to benefit from the long pre-training of the foundational T2A model instead of training all three modalities from scratch. Figure 2a provides a high-level overview of the proposed method.

Text-to-Audio Backbone. A variational autoencoder [9] encodes the input signal $\mathbf{x} \in \mathbf{R}^{2\times L}$ (2 channels for stereo) into a latent representation $\mathbf{z} \in \mathbf{R}^{T\times C}$, with L denoting the temporal length of the audio, while T and the $C = 64$ correspond to the temporal dimension and feature size, respectively. Next, noise is added to the latent representation, producing $\mathbf{z}_t$, which is then processed by a core architecture built from a stack of Diffusion Transformer (DiT) blocks [35], with model-dependent variations. Each DiT block is controlled by a text input, encoded by a pretrained text encoder, guiding the generation process. Furthermore, a timing mechanism sets the signal length and fills the rest with silence.

We experiment with two T2A models: Stable Audio Open [9] and TangoFlux [16]. These models provide the underlying architecture for creating high-quality stereo audio content at a sampling rate of 44.1 kHz based on textual descriptions. Although the architectural designs and sampling methods differ, both models follow a comparable core structure. This allows us to evaluate the flexibility of our approach, demonstrating that it is not tied to a single family of models and can be applied across various T2A models.

Modality Adapter. The adapter is tasked with incorporating the video to guide the T2A model generation. Our design is inspired by ControlNet [54], with some notable differences. First, it operates on the temporal domain, requires synchronization of features from different modalities. Second, our T2A model is a DiT, rather than a U-Net [40], changing the connectivity between the adapter and the base model. Specifically, after the preprocessing stage, informative features from the video, $E_v \in \mathcal{R}^{T\times C}$, are passed through a Zero FC layer and added to $\mathbf{z}_t$. Additionally, the

hidden states, extracted from each DiT block, are processed through Zero FC layers and added to the backbone model, as depicted in Figure 2b. Only the adapter is being trained, while the T2A, Text Encoder, and Video Encoder, all kept frozen.

Video Representation. We experiment with AVCLIP [19] and CLIP [36], both of which are trained with contrastive learning. AVCLIP processes 0.64-second video-audio segments and applies InfoNCE loss [33] to differentiate between positive and negative samples. Its video encoder, built on MotionFormer [34], enhances the modeling of dynamic scenes by capturing implicit motion paths. To ensure better alignment with $\mathbf{z}_t$, we increased the segment overlap compared to the original work, where 216 samples roughly correspond to 10 seconds. We further examine the contribution of CLIP to the semantic scene understanding, similarly to MMAudio [7]. Since CLIP is less sensitive to object locations [3], we introduce a temporal dimension to capture significant visual changes. We extract CLIP representations at 5 FPS and interpolate to match the temporal dimension of $\mathbf{z}_t$. Finally, both representations are preprocessed before being integrated with the modality adapter, as detailed in the supplementary [25].

Asymmetric Classifier-Free Guidance. Incorporating CFG into our model with text as the conditioning signal c presents challenges in effectively utilizing video conditioning at higher guidance scales. To address this limitation, we propose *Asymmetric Classifier-Free Guidance*, where the modality adapter's output is selectively modulated in the conditional and unconditional pathways during synthesis. Unlike the standard approach, which equally integrates the modality adapter into the backbone model in both pathways, our method introduces an asymmetric scaling factor, $0 \leq \alpha \leq 1$, reducing the influence of the adapter in the unconditional path,

$$h'_{i,c} = h_{i,c} \quad , \quad h'_{i,uc} = \alpha \cdot h_{i,uc}, \tag{5}$$

where $h_i = [h_{i,c}, h_{i,uc}]$ are the hidden states, with $h_{i,c}$ representing the conditional path and $h_{i,uc}$ the unconditional path. Consequently, under $\alpha < 1$, this adjustment induced controlled disparity between $\epsilon_\theta(\mathbf{z}_t, t, c)$ and $\epsilon_\theta(\mathbf{z}_t, t)$ effectively amplifies the video conditioning signal. Standard CFG is a special case, corresponding to $\alpha = 1$. Our experiments demonstrate that this simple yet effective modification significantly enhances adherence to video conditioning while maintaining high generation quality and text controllability.

4. Experimental Setup

4.1. Datasets

The proposed model was trained using two datasets: VGGSound [2] and VisualSound [46]. VisualSound is a subset of VGGSound filtered to include samples with high ImageBind [11] scores. Both datasets contain 10-second video clips across diverse acoustic categories accompanied by video captions. For TV2A and V2A evaluation, we use the VGGSound dataset and VGGSound-Sparse [18], which is a subset of VGGSound containing 12 categories of naturally sparse audio events such as "dog barking" or "playing tennis".

4.2. Baseline Methods

We compare CAFA against several state-of-the-art models, namely MMAudio [7] (*large_44k_v2* version), FoleyCrafter [55], VATT [28], ReWaS [20], Frieren [48], and MultiFoley [6]. For FolyCrafter, we follow original configurations by formatting text prompts as "The sound of `<label>`". For VATT, Frieren, and MultiFoley, we consider the samples provided by the respective authors. MultiFoley samples were only available for a subset of the test set with high ImageBind scores. ReWaS is evaluated using its default configuration. Notice, with 5-second samples, unlike other models that produce 8-second samples. Hence, for a fair comparison against this model we truncate the videos to 5-second.

4.3. Implementation Details

CAFA models are initially trained for 48k steps on VGGSound, followed by fine-tuning for 33k more steps on VisualSound. Training was performed with a batch size of 16, using the AdamW [29] optimizer, on a single A100 GPU. We generate samples using CFG = 7, Asymmetric CFG scale of $\alpha = 0.5$, and 50 inference steps, while keeping the rest of the TTA model configuration unchanged. Our model is trained on 10-second samples, and the output truncated to 8 seconds for fair comparison with the baseline methods.

4.4. Evaluation Metrics

We evaluate model performance across four complementary dimensions that capture different aspects of audio-visual generation: Audio Quality, Audio-Visual Semantic Alignment, Audio-Visual Temporal Alignment, and Audio-Text Semantic Alignment.

Audio Quality. We employ three established metrics to assess the fidelity and naturalness of generated audio: (i) Fréchet Audio Distance (FAD) [21] which measures distributional similarity between features extracted from ground truth and generated audio; (ii) Kullback-Leibler Distance (KL) [24], which quantifies the difference between probability distributions of per-sample ground truth and generated audio features; and (iii) Inception Score (IS) [42] that evaluates the generated audio quality independently of ground truth references. We utilize PANNS [22] as the features extractor for all three audio quality metrics.

Model	FAD↓	IS↑	CLAP↑	Acc↑	DeSync↓
FC	57.00	6.10	0.10	0.69	1.30
MMA	**16.43**	6.77	0.10	0.36	**0.57**
ReWaS	38.94	5.13	0.09	0.74	1.19
CAFA (Ours)	27.33	5.63	**0.21**	**0.87**	0.81

Table 1. **Semantically different text and video conditioning.** Our method surpasses strong concurrent SOTA in terms of prompt adherence by a large margin. Arrows indicate whether higher (↑) or lower (↓) values are better. FC: FoleyCrafter, MMA: MMAudio.

Model	FAD↓	KL↓	IS↑	IB↑	DeSync↓	CLAP↑
FC†	13.68	2.56	10.68	0.27	1.30	0.12
MMA†	**5.32**	**1.64**	**17.18**	**0.33**	**0.77**	**0.23**
Frieren†	11.76	2.70	12.33	0.23	1.04	0.11
FC	22.17	2.87	13.30	0.16	1.31	0.18
MMA	**6.89**	1.65	**20.44**	**0.34**	**0.76**	**0.25**
VATT	11.13	**1.48**	11.85	0.25	1.28	0.15
ReWaS*	14.71	2.69	8.45	0.15	1.18	0.18
MF*	13.51	1.65	15.89	0.27	1.04	0.23
CAFA (Ours)	12.60	2.02	13.45	0.21	0.96	0.23

Table 2. **Quantitive comparison.** We report results comparing standard V2A models, V2A variants of TV2A models (indicated by †), TV2A models, and our method. FC:FoleyCrafter, MMA:MMAudio, MF:MultiFoley. $*$ indicates variations - we compare with ReWaS on samples trimmed to 5 seconds, and compare with MultiFoley on their selected subset of the test set.

Audio-Visual Semantic Alignment. We leverage ImageBind (IB) [11] to quantify semantic similarity between the ground truth video and generated audio. This cross-modal embedding model measures whether the generated audio contains appropriate sounds for the visual content.

Audio-Visual Temporal Alignment. We utilize DeSync [7] (also known as Sync [46] or AV-Sync [6]) to measure temporal synchronization between audio and video. DeSync calculates the average absolute offset (in seconds) between ground truth video and generated audio using Synchformer [19] predictions. Following prior work[7], we average DeSync scores from the first and last 4.8 seconds of audio to accommodate Synchformer's limited context window.

Audio-Text Semantic Alignment. We employ CLAP [49] to evaluate similarity between generated audio and textual descriptions of the video by calculating the cosine similarity between the text and audio embeddings.

5. Results

We start by evaluating model performance when considering semantically different text and video conditioning. Ideally, we expect the model to generate textually described audio aligned with the visual cues. For that we generate audio for each video in VGGSound-Sparse, using the ground truth video paired with captions from each of the 11 other categories. This cross-category approach creates a challenging scenario where models must follow textual instructions that deliberately conflict with visual content.

Notice, under this setup we compute the CLAP similarity score between the new caption and generated audio. We also use CLAP as a classifier between the new caption and the GT caption for the generated audio, reporting binary classification accuracy (Acc).

For FoleyCrafter, we follow [6] and disable the semantic adapter to allow the model to generate the requested caption, using only the temporal adapter to retrieve information from the video.

Results are summarized in Table 1, with visual examples depicted in Figure 3. MMAudio achieves the highest audio quality (best FAD and IS scores) and temporal alignment (DeSync), but critically fails at semantic control. Its Acc scores of 0.58 (w\negp) and 0.36 (w\o negp) demonstrate that it systematically generates audio corresponding to the visual content rather than adhering to the requested text prompt, essentially negating the purpose of text-guided generation. Despite disabling FoleyCrafter's semantic adapter specifically for this experiment, it exhibits severe temporal misalignment (worst DeSync score of 1.30) producing outputs that fail to synchronize with visual events. This renders its generation ineffective even for basic audio-visual correspondence. ReWaS performs sub-optimally across all evaluation dimensions—producing lower audio quality, weaker temporal alignment, and poorer semantic control compared to our approach—without exhibiting a particular strength in any area to compensate for these deficiencies. In contrast, CAFA successfully balances all critical requirements: achieving strong audio quality (second-best FAD and comparable IS), effective temporal alignment (second-best DeSync), while significantly outperforming all competitors in semantic controllability with a CLAP score of 0.21 (compared to < 0.1 for all others) and Acc of 0.87. This comprehensive performance profile makes our approach uniquely capable of producing high-fidelity, temporally-aligned audio that accurately follows semantic text instructions.

Semantically Aligned Visual and Textual Conditions. Next, we compare CAFA where the visual and textual conditions are semantically aligned. We compare the proposed method against V2A and VT2A methods, considering the VGGSound test set using the standard configurations. For V2A models, we do not use textual descriptions. Results are presented in Table 2.

While MMAudio emerges as the strongest performer, comparison of MMAudio with and without text conditions reveals minimal benefits from textual inputs. Specifically,

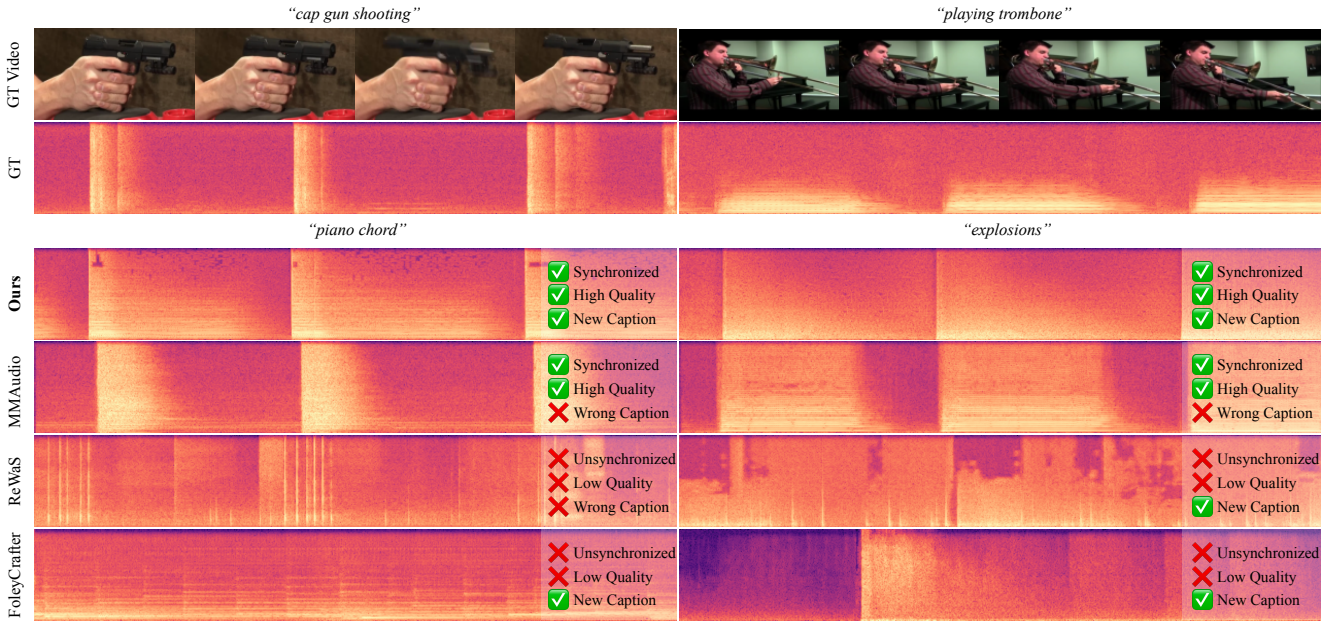

Figure 3. **Qualitative Comparison of Text-Video Disentanglement.** A comparative analysis of various TV2A models: Ground Truth (GT), CAFA (ours), MMAudio, ReWaS, and FoleyCrafter, using the same configurations as in Table 1. Our model consistently delivers synchronized, high-quality generations that accurately adhere to the requested target captions, outperforming other approaches. Full videos presented at *our demo page*.

FAD got slightly worse when text is added, while other metrics remain largely unchanged. Interestingly, a similar pattern emerges in FoleyCrafter, which shows mixed performance with text conditioning - its FAD, KL, IB got worsen, while IS and CLAP improved, and DeSync stays unchanged. These results strengthens our findings in Table 1, demonstrating the improved flexibility and controllability of the proposed method compared to the baseline methods.

CAFA demonstrates balanced performance across all evaluation metrics, achieving the second-best scores in both DeSync and CLAP while maintaining competitive audio quality metrics. These results indicate effective multimodal conditioning without compromising performance in any single dimension.

Human Study. Lastly, we evaluate the subjective quality of audio generated by the proposed method. We conduct a user study using the following evaluation protocol considering both the original textual captions (OC) and visually unaligned text caption (UC). We follow the above mention protocol.

Each participant was presented with a pair of videos side by side. For every pair, the participant was asked three distinct questions: (i) Prompt Adherence (PA): *Which audio better matches the description '[text prompt]'?* This question measured how well the audio corresponds to the prompt; (ii) Alignment: *Which audio better aligns with the timing of visual movements and events in the video?'* This question assessed the synchrony between audio events and the corresponding visual actions; (iii) Quality: *Which audio has higher overall technical quality (considering naturalness, clarity, and lack of artifacts)?* This evaluates the technical fidelity of the audio signal. Participants provided independent responses for each of these questions. In both protocols, the ordering of the video pairs was randomized to mitigate any potential bias. The protocol was implemented using a custom web interface built on Amazon SageMaker Ground Truth. We use 17 videos from VGGSound test set

Comparison	Criterion	%(OC)	%(RC)
CAFA vs MMAudio	Align.	0.24	0.24
	Quali.	0.29	0.26
	PA	0.39	0.87
CAFA vs FoleyCrafter	Align.	0.83	0.95
	Quali.	0.81	0.88
	PA	0.82	0.69
CAFA vs ReWas	Align.	0.96	0.96
	Quali.	0.95	0.89
	PA	0.94	0.77

Table 3. **Human study.** Win rate results (%) of CAFA against three baseline methods. Results are reported for: time alignment (Align.), audio quality (Quali.), and prompt adherence (PA), considering the original caption (OC) and unaligned caption (UC).

Model	Audio Quality			A-V Sem.	A-V Tem.	A-T
	FAD↓	KL↓	IS↑	IB↑	DeSync↓	CLAP↑
CAFA-B	**12.57**	2.04	11.84	0.21	1.00	**0.23**
CAFA	12.60	2.02	13.45	0.21	**0.96**	**0.23**
CAFA-C	14.44	**1.98**	14.18	**0.22**	1.02	**0.23**
CAFA-TF	19.94	2.16	**16.94**	0.20	1.12	**0.23**

Table 4. **Model ablation.** We compare different variants of our model, and argue that our default model CAFA strikes an overall good balance between audio quality, and semantic and temporal alignment to video and prompt. CAFA-B - our model before finetuning on VisualSound. CAFA-C - additionally leverage CLIP visual features. CAFA-TF - use TangoFlux[16] as base T2A model.

to generate two audio samples: with an original caption and a new one by every method. Then each pair of CAFA vs competitor (in random order) was shown to 6 raters and the results were averaged across 6 annotations per pair.

Win rate results of the proposed method against the evaluated baselines are presented in Table 3. When comparing to FoleyCrafter and ReWas the proposed method achieves superior performance across all setups. As expected, when comparing to MMAudio, the proposed method reach inferior performance considering both audio quality and time alignment, however, under the unaligned text and visual conditioning, CAFA reach significantly better performance than MMAudio.

6. Analysis

Model Variations. We first evaluate several architectural variants of CAFA. Specifically, we consider: (i) CAFA-B, which is identical to CAFA but does not include additional finetuning phase on VisualSound, allowing us to evaluate training efficiency; (ii) CAFA-C, which leverages both AV-CLIP and CLIP as visual encoders, combined via MLP. This model was trained for 84k steps on VGGSound train split with a batch size of 8 and the same optimizer as CAFA; and (iii) CAFA-TF, that leverages TangoFlux as base T2A model with AV-CLIP visual encode. CAFA-TF was trained for 32k steps on VGGSound train split with batch size 64 and the same optimizer as CAFA. For variants using StableAudio-Open, we used a CFG value of 7.0, asymmetric CFG $\alpha = 0.5$, and 50 inference steps. For CAFA-TF, we used a CFG value of 4.5, asymmetric CFG $\alpha = 0.8$, and 50 steps.

Results, presented in Table 4, shows that all variants achieve similar performance across metrics, with minimal differences in FAD, IS, KL, and DeSync. The CAFA-C results demonstrate that adding CLIP visual conditioning alongside AV-CLIP provides no meaningful benefits, highlighting that the combination of AV-CLIP with text conditioning is sufficient for effective audio generation.

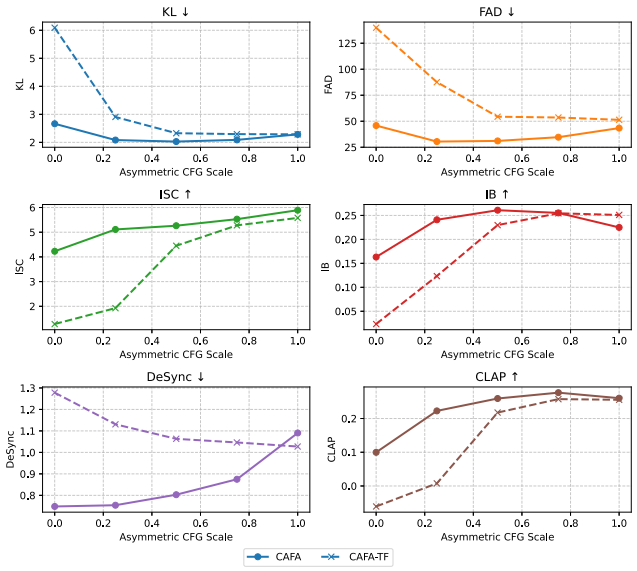

Figure 4. Comparison of Asymmetric CFG Scaling Values. CAFA-TF is our adapter applied with TangoFlux[16], while the default implementation uses StableAudio-Open[9].

The CAFA-B variant, trained for only 48k steps, performs slightly worse than CAFA. Finally, the performance of CAFA-TF confirms that our method generalizes effectively across different T2A base models.

Training Efficiency Analysis. As shown in Table 5 in the Appendix, CAFA demonstrates significant training efficiency compared to other state-of-the-art models. While direct comparison is challenging due to differences in reporting methodologies, hardware configurations, and training approaches, several observations can be made. CAFA requires substantially fewer training steps (81k total) compared to models like Frieren (2.4M steps) and MultiFoley (650k steps). Even our base model (CAFA-B), trained for only 48k steps, achieves performance comparable to models trained for much longer periods. The adapter-based approach allows CAFA to leverage pre-trained text-to-audio models effectively, reducing the need for extensive training from scratch. While MMAudio reports 304 GPU hours on H100 hardware, CAFA estimated 48 GPU hours on A100 hardware represents a more efficient use of compute resources when considering the relative performance. The modular architecture of CAFA facilitates efficient training while maintaining high performance across audio quality, temporal alignment, and textual control metrics. This efficiency analysis further highlights the practical advantages of our approach, making CAFA more accessible for research and potential real-world applications.

Asymmetric CFG Scaling. Finally, we investigate different settings for asymmetric CFG scaling. We evaluate both

CAFA and CAFA-TF over the VGGSound-Sparse test set with the same configurations as described at Section 5. Figure 4 depicts the results, illustrating the trade-off between different scaling parameters. Our analysis shows that for both models, the trends are similar, with values between 0.5 and 1 corresponding to the best results.

7. Related work

Text To Audio Models. The field of Text-to-Audio (T2A) has advanced significantly, with ongoing improvements in text representations and high-quality audio generation. Early approaches included AudioGen [23], an autoregressive model employing a discrete waveform representation, while DiffSound [52] adopted discrete diffusion, removing the need for autoregressive token decoding. As the field evolved, models like AudioLDM [26], StableAudio 1[8], and Make-An-Audio [15] leveraged latent diffusion and incorporated CLAP [49] embeddings to improve text decoding. Recognizing the importance of capturing temporal, acoustic, and semantic information, newer models such as Make-An-Audio 2 [15], AudioLDM 2[27], and Tango [10] integrated large language models (LLMs) [38] to enhance text-audio alignment. Building on these advancements, Tango 2 [32] further refined temporal alignment by employing Direct Preference Optimization (DPO) [37].

Our work is built on StableAudio Open [9] and TangoFlux [16], designed for high-quality text-to-audio generation. StableAudio Open is a latent diffusion model that generates stereo audio up to 47 seconds from text input, utilizing a T5 text encoder for text processing and enabling control over output length. In contrast, TangoFlux is based on rectified flow, producing stereo audio up to 30 seconds at 44.1 kHz, while leveraging a pretrained autoencoder from StableAudio Open to enhance efficiency. Additionally, TangoFlux incorporates CLAP-Ranked Preference Optimization (CRPO) to generate and refine audio preference data.

Video To Audio Models. A key step in automating the Foley process is achieved through Video-To-Audio (V2A) models. Early approaches, such as SpecVQGAN [17], RegNet [4], and FoleyGAN [51], used adversarial training and GAN-based architectures to generate high-quality audio. Diff-Foley [31], a diffusion-based model, introduced CAVP contrastive learning to improve temporal and semantic alignment. Alternatively, Frieren [48], based on rectified flow, enables efficient audio generation in fewer steps. V-AURA [46] adopts an autoregressive approach, leveraging the AVCLIP [19] representation to extract high-framerate temporal and semantic features while bypassing spectrogram conversion.

Beyond models that require training from scratch, V2A-Mapper [47] and Seeing and hearing [50] employ training-free optimization, utilizing pretrained text-to-audio generators or modality mappers to condition audio generation. While these methods reduce computational cost, they often struggle with fine-grained temporal synchronization, highlighting the ongoing challenge of bridging the gap between video and audio in a seamless and efficient manner.

Text and Video To Audio Models. Text-and-Video-to-Audio (TV2A) models introduce text conditioning to enhance control over synthesized audio. VATT [28] leveraged an LLM decoder, functioning as both a video-to-caption model and a video-text-to-audio model. MMAudio [7] and MultiFoley [5] explicitly trained all three modalities from scratch, achieving state-of-the-art results in signal quality and synchronization. While MMAudio introduced a novel network structure for modality fusion, MultiFoley, based on DiT [35], leverages multiple conditioning modalities—text, audio, and video—within a single model. Another approach in TV2A frameworks integrates ControlNet [54] to embed video characteristics into text-to-audio synthesis, as demonstrated by FoleyCrafter [55] and ReWAS [20]. FoleyCrafter extracts frame-based clips as global features in IP-Adapter [53] and trains a timestamp detector to identify sound effect occurrences, integrating this information into ControlNet.

ReWAS, a work closely related to ours, divides the training process into two separate stages: a projection network that learns energy features from video using AVCLIP and a ControlNet that utilizes these features to bridge video and audio. Additionally, AudioLDM serves as the foundation for ControlNet, operating on spectrograms and requiring a vocoder to refine the output audio. In the following sections, we demonstrate that our model surpasses ReWAS in semantic and temporal text-based metrics.

8. Conclusion

In this work, we presented CAFA, a controllable Automatic Foley Artist designed for the video-and-text-to-audio task. Our model ensures high-quality audio synthesis while maintaining both temporal and semantic alignment with the input video. Guided by text prompts, it allows users to incorporate details beyond what is present in the video or even introduce new creative elements. This capability enhances flexibility in sound design beyond video-only Foley models. By leveraging the modality adapter, our approach achieves strong performance on a low computational budget. Both objective metrics and human evaluations confirm its effectiveness in generating high-quality, contextually relevant audio. We believe that further advancements in video feature extraction and T2A model refinement will help address outstanding challenges, such as synthesizing multiple audio sources simultaneously and capturing finer motion details in video.

References

[1] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014.

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset, 2020.

[3] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*, 2024.

[4] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29:8292–8302, 2020.

[5] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. *arXiv preprint arXiv:2411.17698*, 2024.

[6] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. 2025.

[7] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*, 2024.

[8] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*, 2024.

[9] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

[10] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598, 2023.

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[15] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

[16] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.

[17] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

[18] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Sparse in space and time: Audio-visual synchronisation with trainable selectors. *arXiv preprint arXiv:2210.07055*, 2022.

[19] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE, 2024.

[20] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024.

[21] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

[22] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition, 2020.

[23] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[24] Solomon Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[25] FirstName LastName. Supplemental material - cafa: a controllable automatic foley artist, 2025. Supplied as supplemental material `supplemental material.pdf`.

[26] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[27] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[28] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see–video to audio generation through text. *arXiv preprint arXiv:2411.05679*, 2024.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[31] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:48855–48876, 2023.

[32] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572, 2024.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[41] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[42] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *ArXiv*, abs/1606.03498, 2016.

[43] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[46] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. *arXiv preprint arXiv:2409.13689*, 2024.

[47] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15492–15501, 2024.

[48] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37:128118–128138, 2025.

[49] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[50] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161, 2024.

[51] Manjie Xu, Chenxing Li, Xinyi Tu, Yong Ren, Rilin Chen, Yu Gu, Wei Liang, and Dong Yu. Video-to-audio generation with hidden alignment. *arXiv preprint arXiv:2407.07464*, 2024.

[52] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

[53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[55] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*, 2024.